Some notes to clarify Omitted Variable Bias example on aspirin intake and heart attacks

Towards the end of our class on Monday Feb 5th we used an example to identify the sign of an omitted variables bias (OVB), a case dealing with aspirin intake (treatment, or X of interest) and heart attacks, before 65 years-old (the omitted variable) and after 65 years-old (outcome variable). Given that there was a fair amount of confusion as the class approached its end and that the group (including me) have little or no prior idea on the relationship between aspirins (treatment) and heart attacks before 65 years old (omitted variable), I am writing this note to make sure the most important concepts and the sign of the bias itself are well understood.

The exact formulation of the example was at follows:

"For 65 year-olds, each additional daily dose of aspirin is associated with 0.2 fewer heart attacks per lifetime. Describe the bias from omitting the number of heart attacks prior to age 65 as an explanatory variable".

The questions formulated and their corresponding answers follow:

1) Write down the short and long regressions

Long regression: $\#HApost\_65 = \hat{\beta}_0 + \hat{\beta}_1 * aspirin + \hat{\beta}_2 * \#HApre\_65 + \widehat{u}$

Short regression: $\#HApost\_65 = \hat{\alpha}_0 + \hat{\alpha}_1 * aspirin + \widehat{v}$

2) Determine the (expected) signs of $\widehat{\boldsymbol{\beta}}_2, \widehat{\boldsymbol{\gamma}}_1$ and the bias

Here the class seemed to agree that the correlation between heart attacks before 65 years-old (#HApre_65) and heart attacks after 65 years-old (#HApost_65) was **positive.**

$$Corr(\#HApre\_65; \#HApost\_65) > 0$$

The different interpretations came when it came to put a sign to the correlation between heart attacks before 65 (#HApre_65) and aspirin intake (aspirin). As we solved the problem in class – as portrayed in the PDF version of the slides with solutions I uploaded to the Class Page in CANVAS – we assumed that correlation was **positive**. Namely, when people suffered heart attacks before 65 years old, their doctors prescribed taking aspirins to somewhat alleviate their condition.

Some people in the class thought that taking aspirins before 65, would actually help in having less heart attacks, rendering that correlation **negative**. Here I want to provide a solution to the last two questions for both cases, as within the context of this class we do not care much if you know about aspirins and heart attacks, but rather if under certain assumptions you are able to guess the sign of OVB.

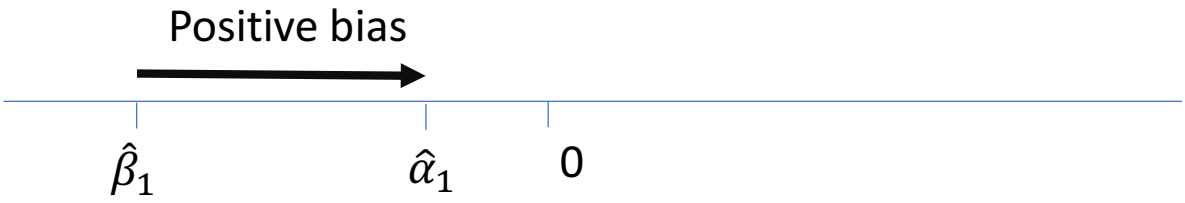
Case 1: Corr(#HApre_65; aspirin) > 0

3) Explain the sign of the bias in words

In this case, we have:

- Corr(#HApre_65; #HApost_65) > 0
- Corr(#HApre_65; aspirin) > 0
- Bias > 0

Given that both correlations are positive, we would expect the bias to be positive. If we locate both coefficients from the short ($\hat{\alpha}_1$) and long ($\hat{\beta}_1$) regression in a line, we might be able to better see what does mean that the *bias is positive* in this case.



Actually, $\hat{\alpha}_1$ is higher than its true estimated impact $\hat{\beta}_1$ (in the negative scale, closer to zero). In the context of this particular example, that implies that aspirins have a much higher negative impact than we will be able to assess if we omit #HApre_65.

4) Determine whether $\hat{\alpha}_1$ is an under- or over-estimate of the true impact of the "treatment"

The intuition here goes as follows: *Aspirin* has a negative relationship with *#HApost_65* (that you can infer from the word *fewer*). At the same time, #HApre_65 is positively correlated with both, #HApost_65 and aspirin, so when #HApre_65 moves, it causes these two variables to move *in the same* direction. If we do not take into account #PApre_65, the effectiveness of aspirins in reducing heart attacks post 65 will be underestimated, as $\hat{\alpha}_1$ will be higher (or less negative) than $\hat{\beta}_1$.

I apologize for the confusion in class, which might have resulted from the fact that in this particular example, the OVB causes the coefficient to be higher in the short regression ($\hat{\alpha}_1$ is less negative than $\hat{\beta}_1$); but the actual effect of treatment is underestimated (aspirins are more effective in reducing heart attacks than our short regression indicates).

To avoid confusion, **we will only refer to underestimation or overestimation when we talk about the impacts**, and **not when we talk about the coefficients**. In this particular case, the impact of treatment (aspirin intake to reduce heart attacks) is underestimated.

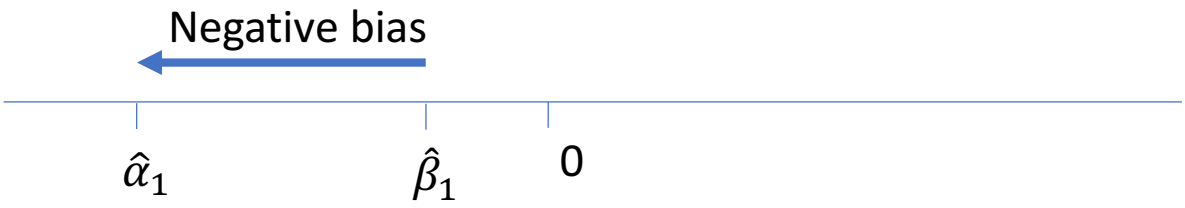
Case 2: Corr(#HApre_65; aspirin) < 0

3) Explain the sign of the bias in words

In this case, we have:

- Corr(#HApre_65; #HApost_65) > 0
- Corr(#HApre_65; aspirin) < 0
- Bias < 0

Given that both correlations have opposite signs, we would expect the bias to be negative. If we locate both coefficients from the short ($\hat{\alpha}_1$) and long ($\hat{\beta}_1$) regression in a line, we might be able to better see what does mean that the *bias is negative* in this case.



Actually, $\hat{\alpha}_1$ is lower (more negative) than its true impact $\hat{\beta}_1$ (in the negative scale, further away from zero is lower). In the context of this particular example, that implies that aspirins have a much lower negative impact than we will be able to assess if we omit #HApre_65.

4) Determine whether $\hat{\alpha}_1$ is an under- or over-estimate of the true impact of the "treatment"

The intuition here goes as follows: *Aspirin* have a negative relationship with *#HApost_65* (that you can infer from the word *fewer*). The problem here is that #HApre_65 is negatively correlated to aspirin, and positively correlated to #HAposr_65. When #HApre_65 moves, it causes these two variables to move *in opposite* directions. If we do not take into account #PApre_65, the effectiveness of aspirins in reducing heart attacks post 65 will be overestimated, as $\hat{\alpha}_1$ will be lower (or more negative) than $\hat{\beta}_1$.

Here again, you might be confused by the fact that while in this case the OVB causes our estimate in the short regression to be lower ($\hat{\alpha}_1$ will be lower or more negative than $\hat{\beta}_1$); the actual effect of treatment is overestimated (aspirins are less effective in reducing heart attacks than our short regression indicates).

Following the same guideline as above – using overestimation or underestimation only as related to treatment effects – in this case the impact of treatment (aspirin intake to reduce heart attacks) is overestimated.